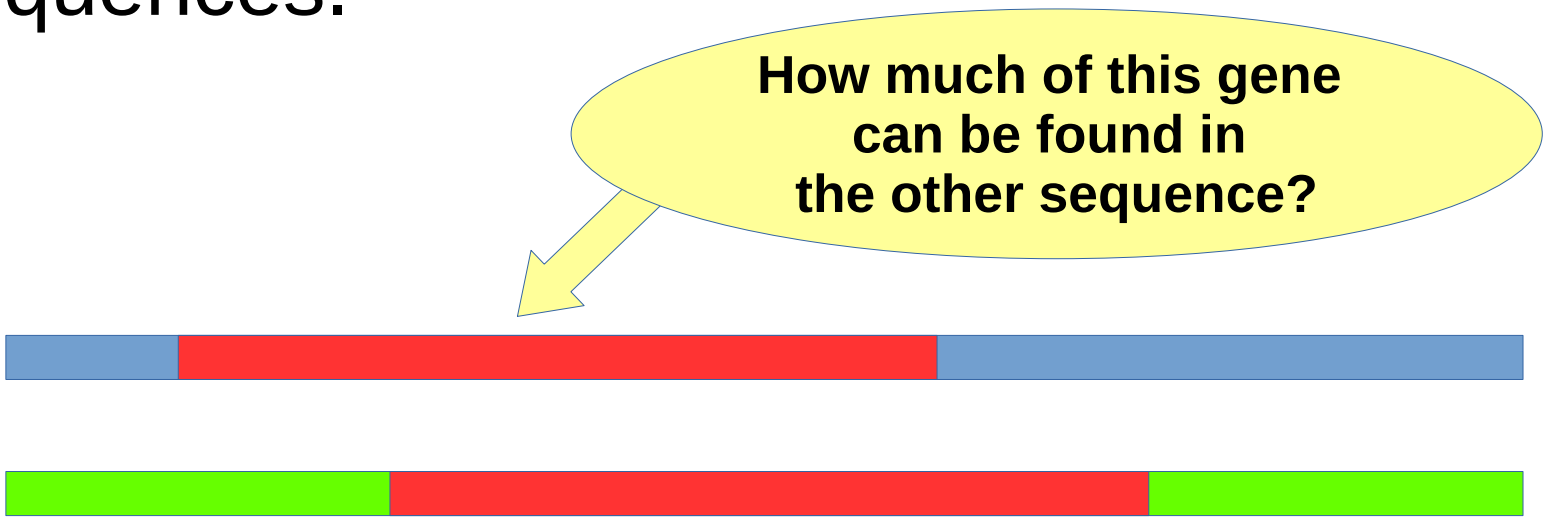
Bioinformatics
CS300
Chap 3
Sequence Alignment with ClustalW

Week 4, Deck 2
Fall 2022
Oliver BONHAM-CARTER

What is Sequence Alignment?

- Sequence alignment is a way of arranging the sequence of genetic material (DNA, RNA or protein) to identify regions of similarity that may be a consequence of functional, structural or evolutional relationships between the sequences.



Types of Alignment

A	T	C	G	A	A	C	T	G	G	C	C	–	–
.	.	|	|	.	|	|	|			|	|		
T	A	C	G	C	A	C	T	–	–	C	C	A	A

(a) Global alignment example

–	A	T	C	G	A	A	C	T	G	G	C	C	–	–
			|	|	.	|	|	|			|	|		
T	A	–	C	G	C	A	C	T	–	–	C	C	A	A

(b) Semi-global alignment example

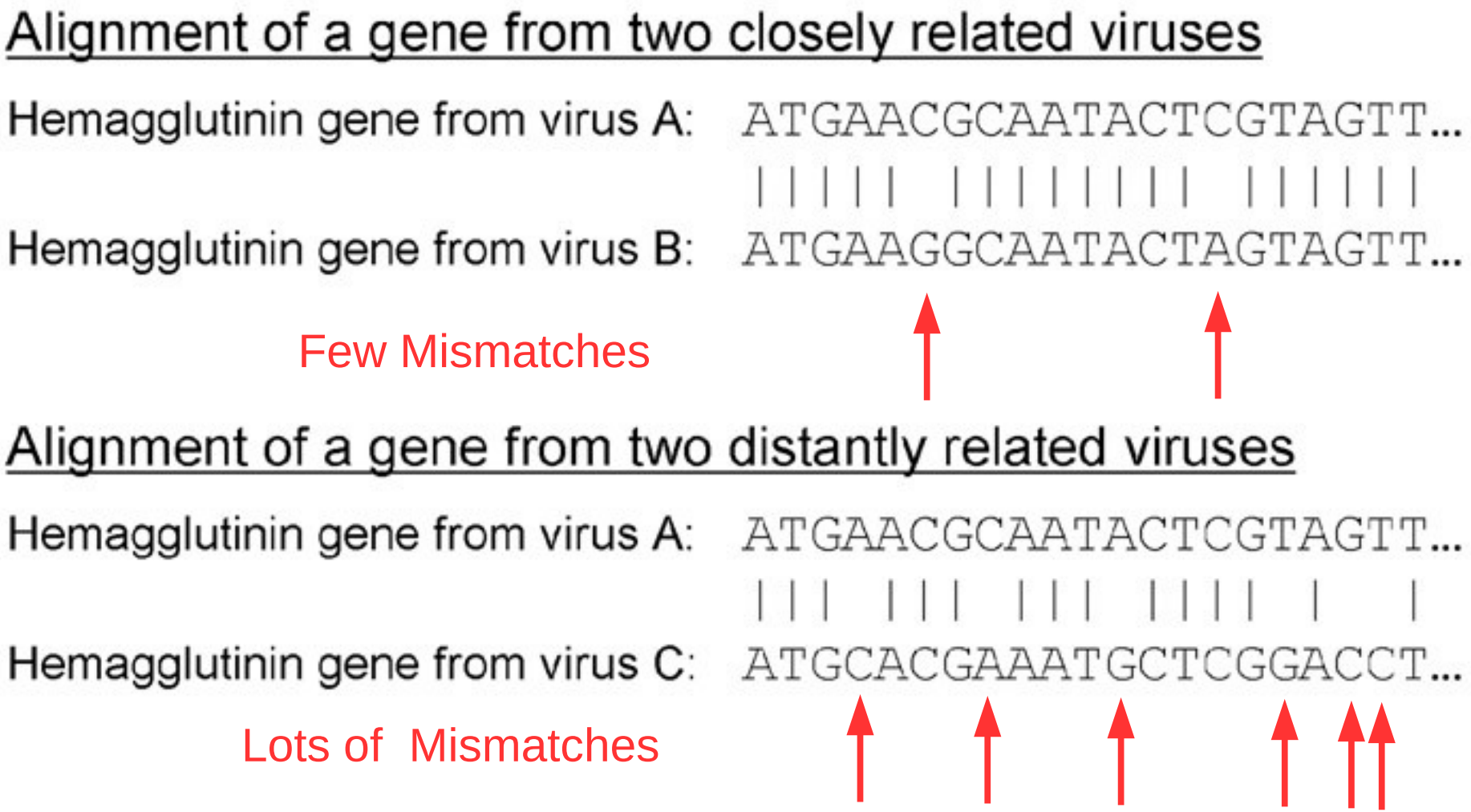
A	T	C	G	A	A	C	T	G	G	C	C		
		|	|	.	|	|	|			|	|		
T	A	C	G	C	A	C	T	–	–	C	C	A	A

(c) Local alignment example

Needleman-Wunsch Algorithm Background

- **Alignment**: Used to determine which parts of a sequence are in common with another sequence; used to measurement similarity between sequences.
- Developed by Saul B. Needleman and Christian D. Wunsch in 1970.
- Dynamic programming to find optimal solution for matching the characters of the two sequences.

Ex: Pairwise Alignment



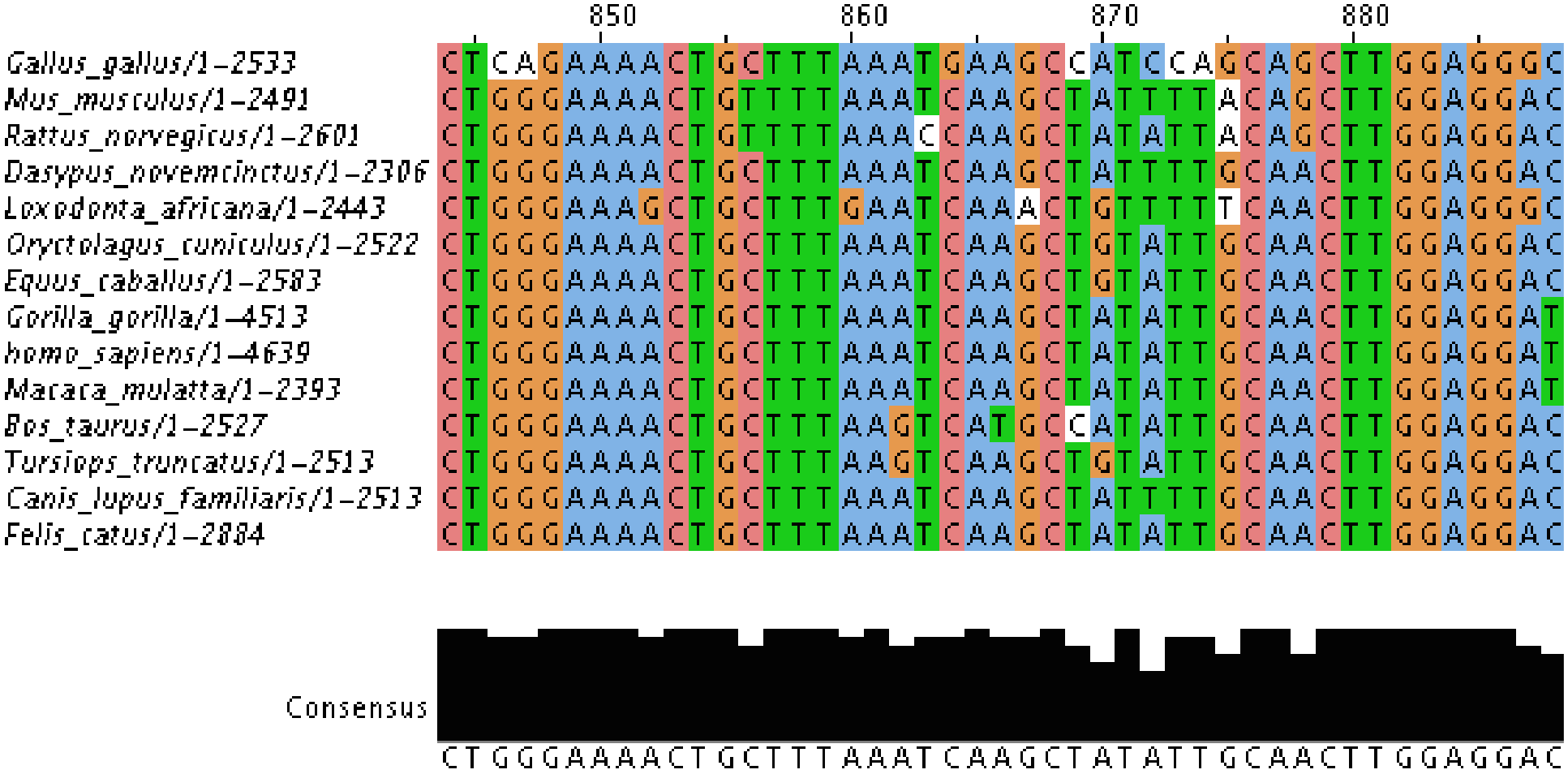
What is Global Sequence Alignment?

- We search for matches, matches and gaps between two sequences to determine their **relatedness**.
- (*) indicate matches or similar nucleotides (bases) along sequence
- Here, the sequences may have a common ancestor

```
ACGTACT    ACGTAC-T    ACGTACT----
ACTACGT    AC-TACGT    ----ACTACGT
**    *    ** *** *        ***
```

Ex: Comparing DNA

- We compare DNA samples from several different organisms.

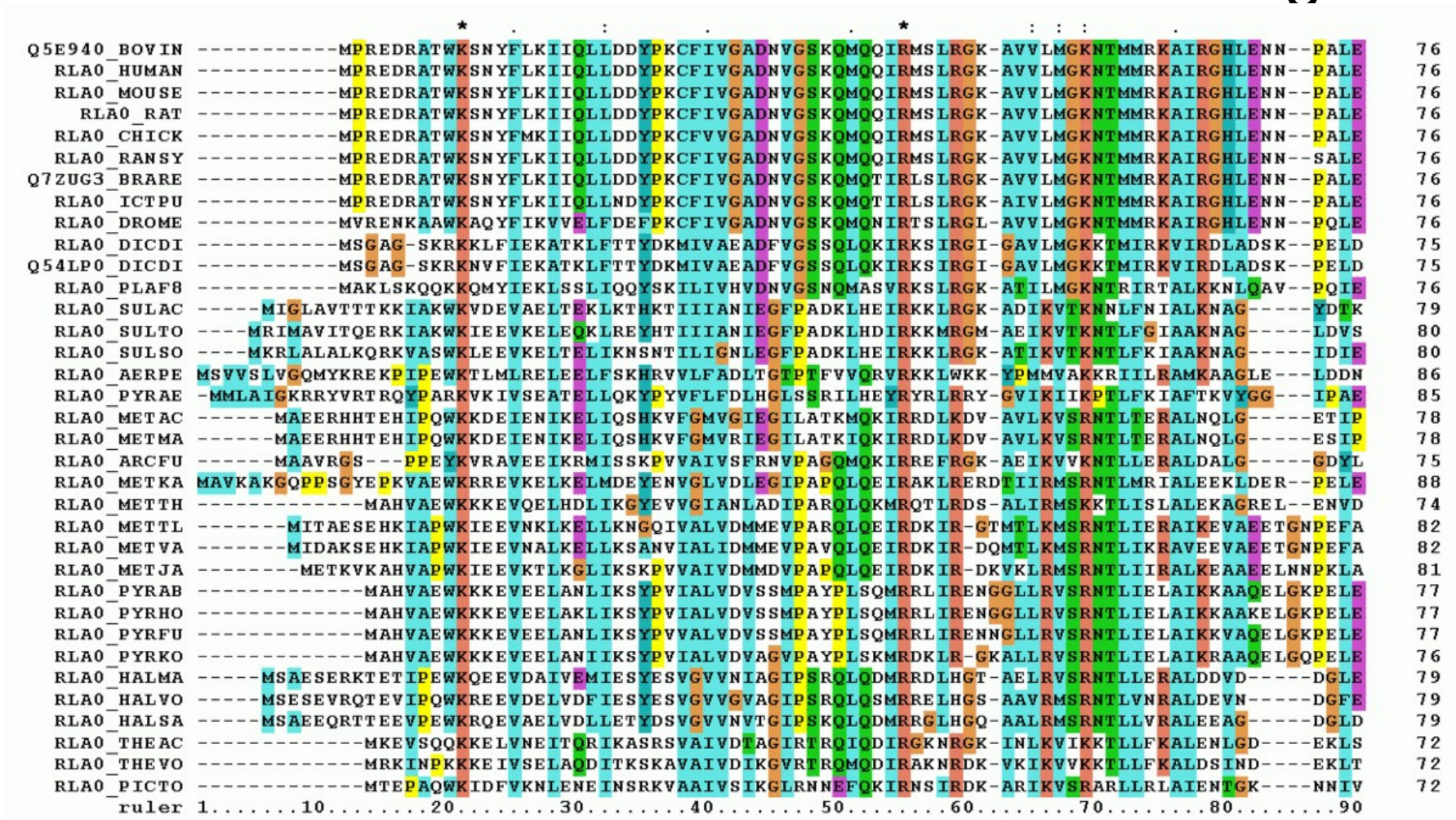


Ex: Comparing Protein

- We compare protein samples from several different organisms.

Ex: Comparing Protein

- The common code have same nucleotides but there are still breaks in these common regions.



Terms

- Alignment is divided up unto sub problems
- Solutions are scored; the best solutions for char by char comparison are kept in the overall solution.
- **Match** – bases of each sequence at position ARE same
- **Mismatch** – bases of each sequence at position are NOT same
- **Gap** – bases are not the same, some insertion or deletion may have occurred.

AGGCTATCACCTGACCTCCAGGCCGATGCCC
TAGCTATCACGACCGCGGTCGATTTGCCCGAC

-AGGCTATCACCTGACCTCCAGGCCGA--TGCCC---
TAG-CTATCAC--GACCGC--GGTCGATTTGCCCGAC

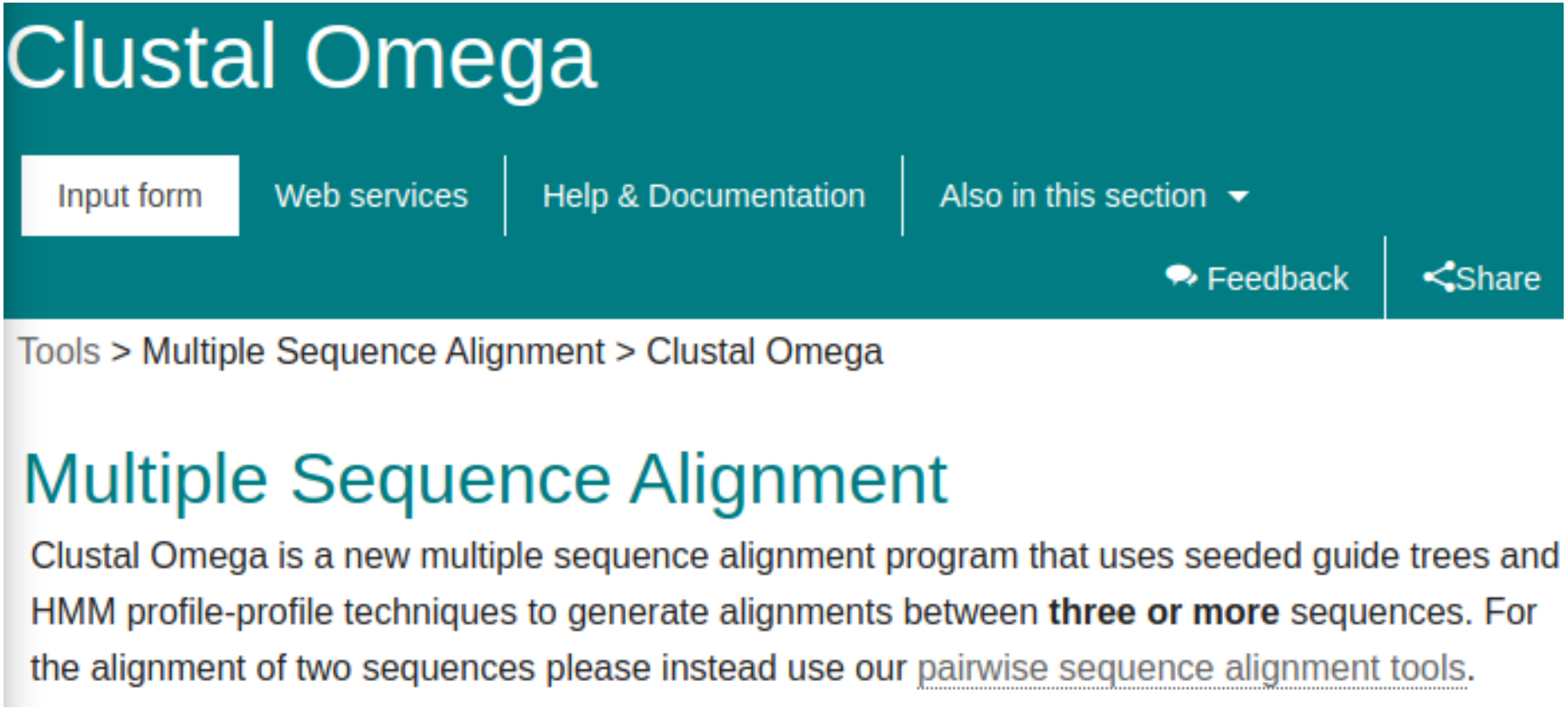
Terms

- **Homology** – Two or more sequences have a common ancestor
- **Similarity** – Two sequences are similar in terms of base arrangements. Note: this similarity does not refer to any specific evolutionary process; the sequences show *similarity* as they are compared.
- **Conserved regions** – Regions in code which are very similar (or the same) across a wide group of organisms. Having code which has not changed, in light of mutations, in all the organisms suggests that the region have been maintained by natural selection (and may serve an important function.)
- **DNA Coding Regions** – Contains code that is more likely to make protein, often less likely to change genetically. Mutations in these areas may cause danger.
- **DNA NonCoding Regions** – Contains DNA that does not necessarily code for protein, but may serve in gene regulation, such as the binding or recognition sites of ribosomes and transcription factors. May still be be conserved within a genome.

Bring the Tool!

Up Next!

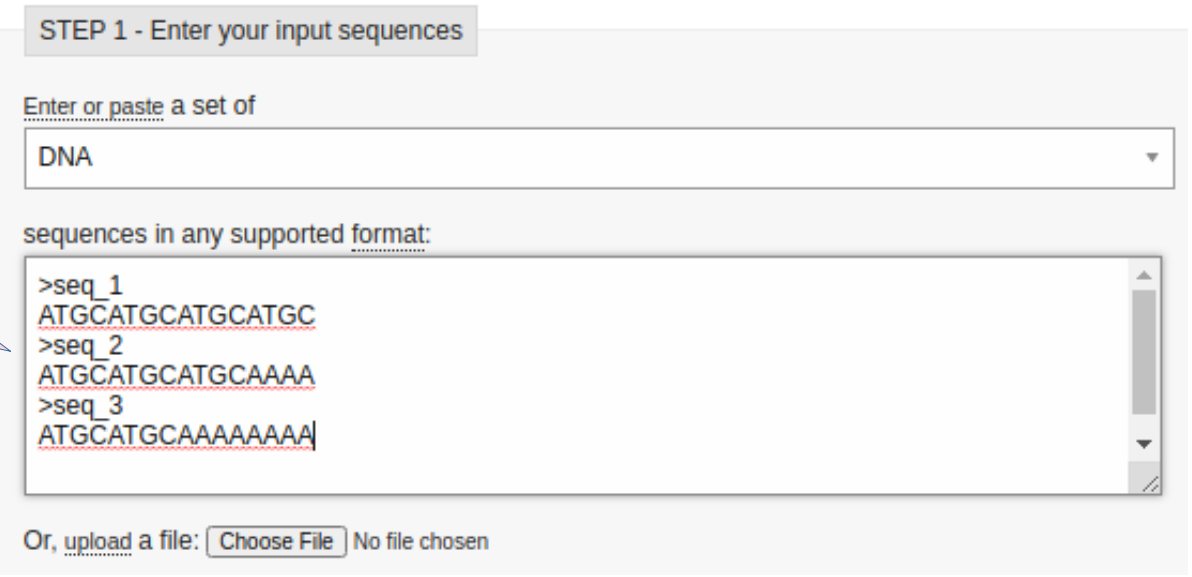
Clustal Omega
Multiple Sequence Alignment

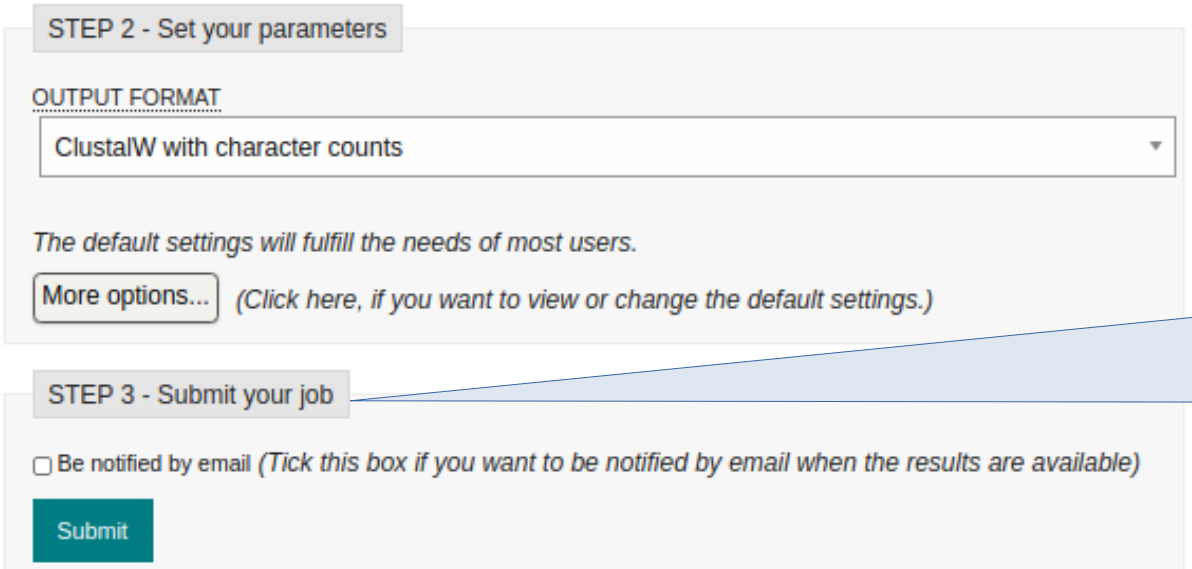


Link:
https://www.ebi.ac.uk/Tools/msa/clustalo/

Clustal Omega
Multiple Sequence Alignment

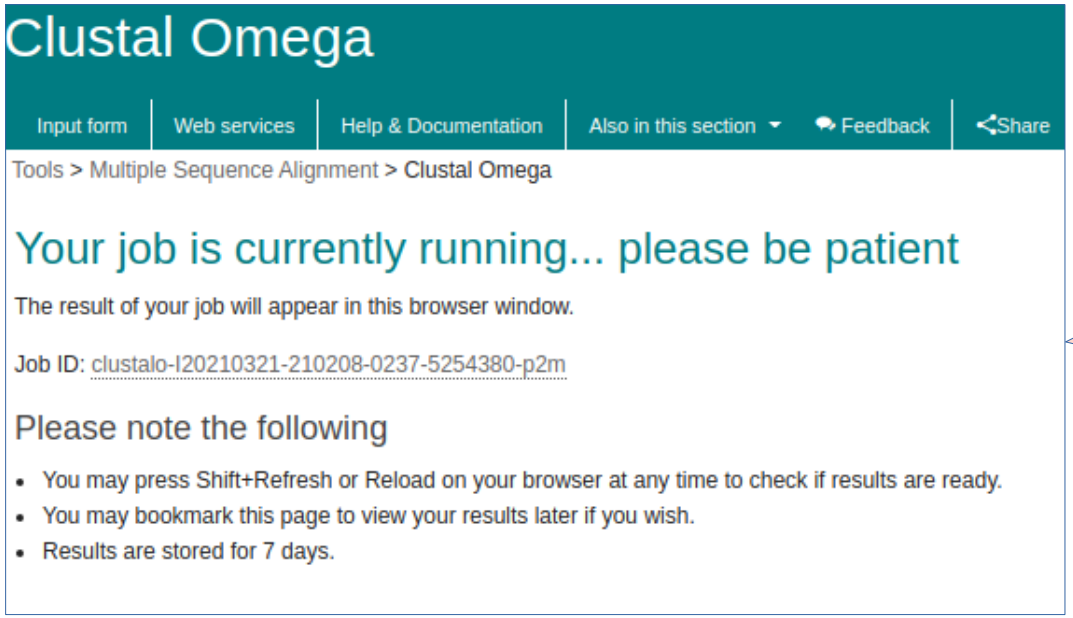
Enter DNA sequences in this **FASTA** format.





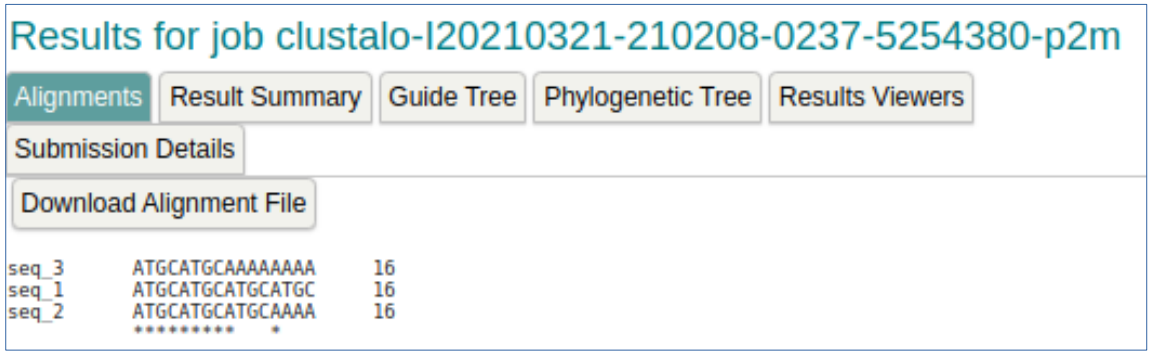
Submit your job with the *Character Counts* option.

Clustal Omega Results

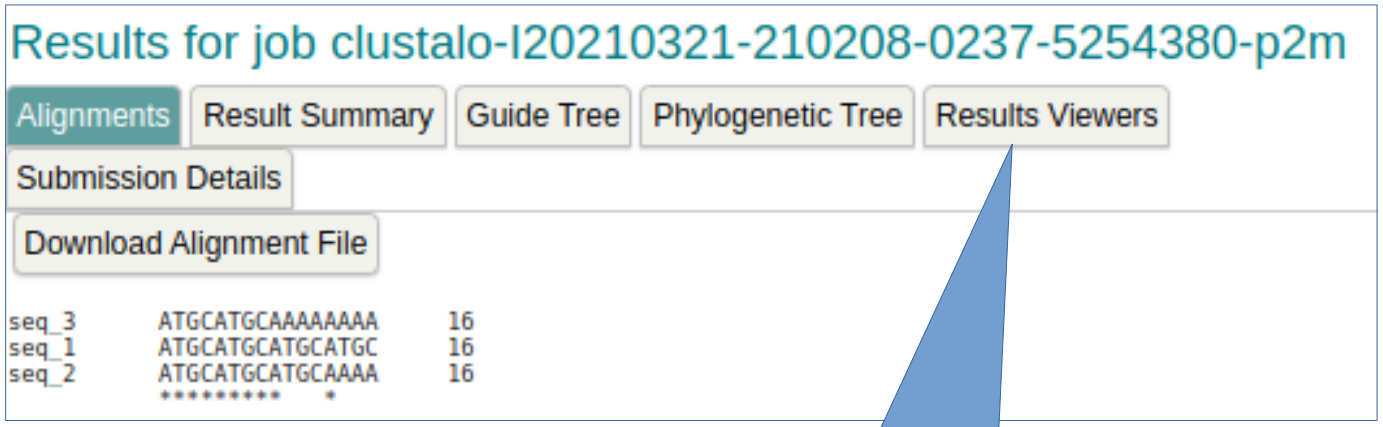


Depending the number of seqs, you may have to wait...

The '*'s denote agreement in bases across sequences.



Percent Identity Matrix
How similar are the seqs?

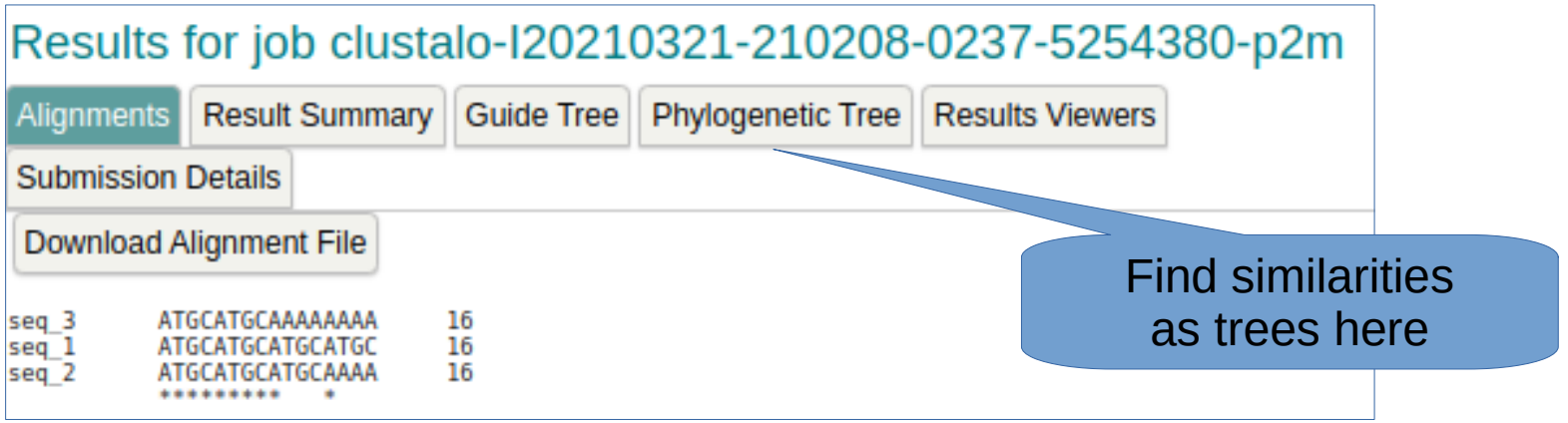


How similar are the sequences?.

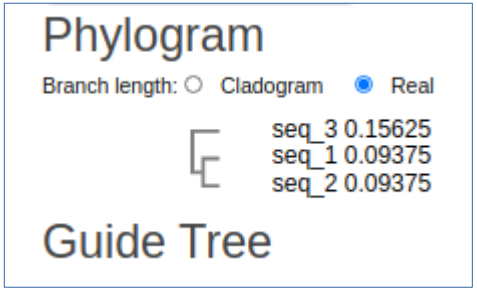
```
#
#
#  Percent Identity  Matrix - created by Clustal2.1
#
#                        Seq_3  Seq_1  Seq 2
     1: seq_3          100.00   62.50   81.25
     2: seq_1           62.50  100.00   81.25
     3: seq_2           81.25   81.25  100.00
```

Percent Identity Matrix

How similar are the seqs?

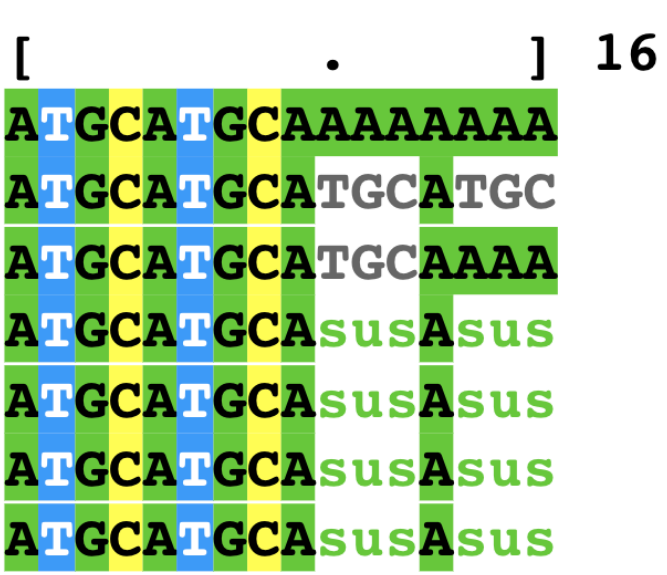


View similarity in a tree from.



Alignment Matrix
PID

		cov	pid
1	seq_3	100.0%	100.0%
2	seq_1	100.0%	62.5%
3	seq_2	100.0%	81.2%
	consensus/100%		
	consensus/90%		
	consensus/80%		
	consensus/70%		

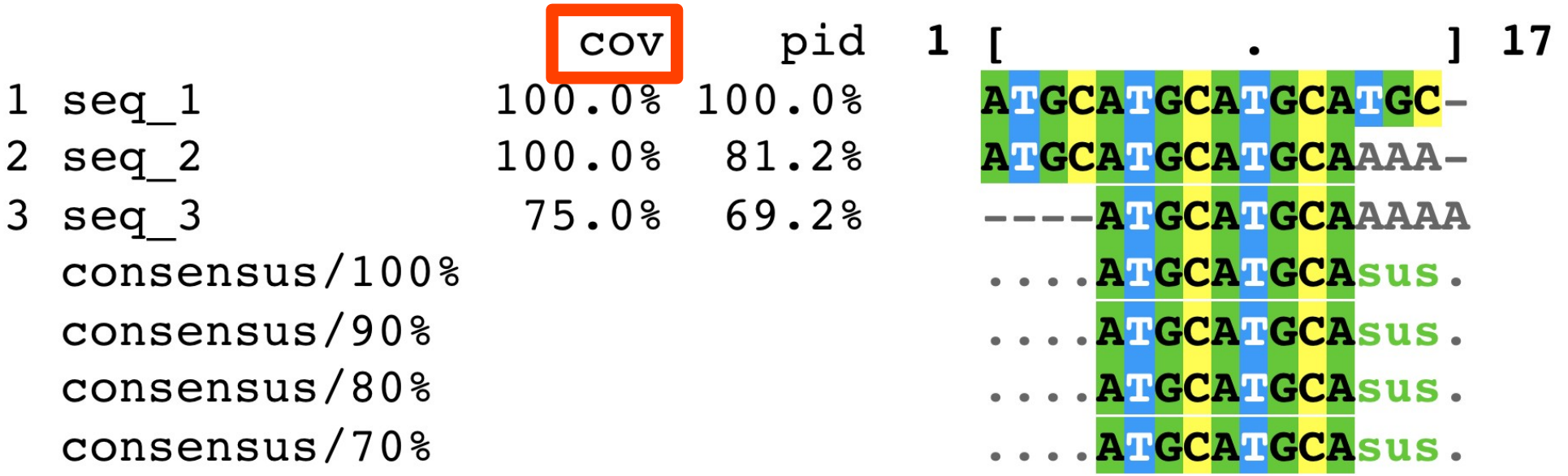


What does percentage identity (PID) refer to?

- This value is a single numeric score determined for each pair of aligned sequences.
- It measures the number of identical residues (“matches”) in relation to the length of the alignment.

- Compare each sequence to the top one.
  - Seq3 vs Seq1: **10** Matches, 6 Mismatches, Length = **16**, Pid = **10/16** = 0.63
  - Seq3 vs Seq2: **13** Matches, 3 Mismatches, Length = **16**, Pid = **13/16** = 0.82

Alignment Matrix
COV



What does coverage (COV) percent refer to?

- This value is a ratio of the length of one sequence against the length of another.
- It measures the number of potential pairing bases.

- Compare each sequence to the top one.
  - Seq3 vs Seq1: **Len(Seq3)** / **Len(Seq1)** = **12** / **16** = 0.75

ClustalW: For Local Use

```
Preparing to unpack .../clustalw_2.1+lgpl-6build1_amd64.deb ...
Unpacking clustalw (2.1+lgpl-6build1) ...
Setting up clustalw (2.1+lgpl-6build1) ...
Processing triggers for man-db (2.9.1-1) ...
obonhamcarter@jupyter-cs-allegheny-edu:~$ clustalw



 ********************************************************************
 ******** CLUSTAL 2.1 Multiple Sequence Alignments  ********
 ********************************************************************


     1. Sequence Input From Disc
     2. Multiple Alignments
     3. Profile / Structure Alignments
     4. Phylogenetic trees

     S. Execute a system command
     H. HELP
     X. EXIT (leave program)


Your choice: x
obonhamcarter@jupyter-cs-allegheny-edu:~$
```

http://www.clustal.org/clustal2/#Download

ClustalW: For Local Use

Enter sequences in a FASTA format:

```
>seq_1
ATGCATGCATGCATGC
>seq_2
ATGCATGCATGCAAAA
>seq_3
ATGCATGCAAAAAAAA
```

```
root@ea16b8965382:~# clustalw samples.fasta


 CLUSTAL 2.1 Multiple Sequence Alignments


Sequence format is Pearson
Sequence 1: seq_1         16 bp
Sequence 2: seq_2         16 bp
Sequence 3: seq_3         16 bp
Start of Pairwise alignments
Aligning...

Sequences (1:2) Aligned. Score:  81
Sequences (1:3) Aligned. Score:  56
Sequences (2:3) Aligned. Score:  75
Guide tree file created:   [samples.dnd]

There are 2 groups
Start of Multiple Alignment

Aligning...
Group 1: Sequences:   2      Score:256
Group 2: Sequences:   3      Score:214
Alignment Score 236

CLUSTAL-Alignment file created  [samples.aln]
```